Aalto University
School of Business

SENTIMENT ANALYSIS IN SALES ESTIMATION

An Econometric Analysis of Product Listings and Reviews in a Chinese Cross-Border E-Commerce Context

Bachelor's Thesis
Visa Mäkeläinen
Aalto University School of Business
Information and Service Management
Spring 2021

Aalto University
School of Business

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi

Abstract of Bachelor's thesis

Author Mäkeläinen, Visa

Title of thesis Sentiment Analysis in Sales Estimation: An Econometric Analysis of Product Listings and Reviews in a Chinese Cross-Border E-Commerce Context

Degree Bachelor's degree

Degree programme Information and Service Management

Thesis advisor(s) Seppälä, Tomi

Year of approval 2021 **Number of pages** 23 **Language** English

Abstract

Since the advent of electronic word-of-mouth communication, particularly in the form of user-generated reviews on e-commerce platforms, research has been undertaken to quantify and draw insights from this growing wealth of data. Coinciding developments in machine learning and natural language processing have enabled the systemic analysis of these texts, heightening the role of user feedback from a simple information channel between users to an indispensable source of "big data" information regarding consumer sentiment and behaviour.

While otherwise extensive, contemporary research into the role of consumer sentiment, and, in particular, its effect on sales outcomes, is largely built around data gathered from Western e-commerce platforms, most notably Amazon. This has potentially limited its generalizability to wider contexts. In addition, many studies simplify the role of feedback valence by interpreting the sentiment polarity of a written review as equivalent to its corresponding numerical rating – a conflation that seems to go against existing research into rating inflation and other biases.

This study seeks to further this field of e-commerce research by accounting for these issues. Using cross-sectional data gathered from an industry-leading Chinese cross-border e-commerce platform, this study analyses the relationships between user-generated review sentiments and order amounts in a new context. By applying three different sentiment analysis tools to a total of 451,375 product reviews, overall sentiment polarity and subjectivity metrics were calculated for 8,319 product listings. Using these values, alongside other control variables (including numerical ratings, separate from sentiment polarities) from the listings, econometric regression models describing the relationships were estimated and interpreted.

The findings of this study demonstrate that, on a broad level, the notion of review sentiment polarity being positively related to sales outcomes is generalizable beyond the Western context. The role of a more nuanced aspect of review sentiments, namely the subjectivity of reviews, is found to be seemingly different from existing research into Western platforms, albeit somewhat inconclusively. The findings also support the notion that review sentiment polarity is not directly represented by its corresponding numerical rating, and that future studies should continue to differentiate between these two metrics.

This study leaves open the exact causal nature of these relationships, requiring future research using time series data over multiple years. In addition, a greater variety of product categories could be studied in order to confirm the overall generalizability of these findings.

Keywords sentiment analysis, big data, e-commerce, online reviews, business analytics

Table of Contents

Abstract
1 Introduction 1
1.1 Research objectives and research questions 2
1.2 Scope of research 3
1.3 Structure of the research 4
2 Theoretical background 5
2.1 Reviews, ratings, and sales 5
2.2 Sentiment analysis in sales estimation 6
2.3 Chinese cross-border e-commerce 7
3 Methodology 9
3.1 Data collection procedure 9
3.1.1 Ethics of data collection 9
3.2 Data preparation and variables 10
3.2.1 Listing variables 11
3.2.2 Pricing variables 11
3.2.3 Review variables 14
3.3 Empirical econometric modelling 16
4 Results 17
5 Discussion and conclusions 19
5.1 Implications to research 20
5.2 Implications to practice 21
5.3 Limitations and future research 22
References 24
Appendices 30

List of Figures

Figure 1. System design of the data preparation process 10
Figure 2. Example listing with outlier price options 12

List of Tables

Table 1. Definitions of variables 16
Table 2. Summary statistics 16
Table 3. OLS regression estimates 17
Table 4. Multicollinearity-corrected OLS regression estimates with β_4 *sale* excluded 18

1 Introduction

2021 marked the estimated fiftieth anniversary of what has been considered the first-ever e-commerce transaction, taking place between Stanford students on ARPANET in 1971 or 1972 (Markoff, 2006). Since breaking out into the mainstream in the 1990s, e-commerce has become an integral part of all commerce, and providers such as Amazon and Alibaba Group continue to expand over the former spheres of physical retailers. A driving force in the success of these platforms has been the role of user-submitted product evaluations, consisting of both numerical ratings (e.g. stars on a discrete 1 to 5 scale) and plaintext written reviews. By encouraging electronic word-of-mouth (eWOM) information spread, e-commerce platforms have synergistically both provided additional information channels to mitigate perceived risk (Dellarocas, 2003) and created an indispensably valuable source of "big data" – for their own analysis, and others', as well.

The ever-increasing popularity of e-commerce has coincided with developments in natural language processing (NLP), or the computerized analysis of human language using statistical machine learning and deep learning methodologies. Sentiment analysis, the subfield of NLP dealing with affective states and subjective statements, has been extensively applied in systematically analysing written user reviews (Pang & Lee, 2008). Common approaches include quantifying user-generated text into "polarity" scores (the positivity or negativity of the sentiment) and "subjectivity" scores (the degree of opinion in the text, as opposed to objective fact) (Pang & Lee, 2004). Among other uses, these figures have been utilized in econometric regression models as explanatory variables, in order to estimate, for example, sales outcomes (Ghose & Ipeirotis, 2006; Hu *et al.*, 2014).

Contemporary study into the role of user reviews in e-commerce product sales has been largely focused on data from Western platforms, particularly Amazon. However, developments in information technology, international shipping, and governmental policies have increasingly enabled expansion in cross-border e-commerce (CBEC), particularly from cost-competitive countries, such as China (Liu & Liu, 2017; Wang *et al.*, 2017). Alibaba Group's Chinese cross-border B2C platform, AliExpress, has seen significant growth in recent years, becoming a leading global e-commerce website and one of Amazon's primary competitors (Xu, 2016; Lukicheva & Semenovich, 2019). Despite this, compared to Amazon, AliExpress has remained largely underrepresented in these studies. To address this imbalance in research, this study seeks to analyse the relationship between review sentiments and product orders, using AliExpress product listings and their reviews.

In addition, a potential oversight in several of the aforementioned high-profile studies has been the equation of user-submitted numerical ratings with genuine review sentiment. Many published papers, including those presenting new NLP training methodologies, assume that the numerical rating a user submits corresponds to the positivity/negativity of their written review (Pang *et al.*, 2002; Ghose & Ipeirotis, 2007). However, recent research has shown that user-submitted ratings are heavily influenced by factors such as rating inflation (Filippas *et al.*, 2018; Garg & Johari, 2020; Lee, 2020), leading to considerations about the accuracy of numerical ratings as a whole (Hu *et al.*, 2009; Lee, 2020). To explore this potential disconnect further, this study uses both user-submitted numerical ratings and NLP-derived sentiment scores to estimate sales outcomes.

1.1 Research objectives and research questions

The objectives of this study are twofold:

1. To further this field of research by studying the impact of user sentiment in an alternative e-commerce context, namely Amazon's Chinese competitor, AliExpress.
2. To achieve a more nuanced understanding of user sentiment impact by differentiating the effects of textual review sentiment from numerical ratings using general-purpose sentiment analysis models.

These objectives can be written out as the following research questions:

1. To what extent are current theories regarding the relationships between user sentiment and sales applicable in a Chinese CBEC context?
2. Can a discrepancy between the effects of user-submitted numerical ratings and NLP-derived review polarities be identified using reviews and listings from a Chinese CBEC platform?

In essence, this study intends to formulate cross-sectional econometric models using empirical data gathered from the AliExpress website. These models examine the role of derived sentiment metrics, specifically polarity and subjectivity scores, as well as average numerical ratings in estimating sales figures (order amounts). Control variables include the price and the age of the given listing, among others. The sentiment metrics are calculated from the product listings' user-generated reviews, using three leading open-source NLP libraries for the Python programming language.

This study is novel, as it is among the first to utilize listing and review data from the scope of Chinese cross-border e-commerce. It investigates the key differences between this context and more commonly studied Western e-commerce platforms, particularly Amazon. An additional advantage of using AliExpress as a source (as opposed to Amazon) is that it reports recent[1] order amounts for each product listing, giving a more accurate approximation of sales outcomes than the "Sales Rank" (Floyd *et al.*, 2014) figures reported by Amazon and used as a proxy by many other studies.

This study is also among the first to utilize the relatively recently developed, "state-of-the-art" NLP framework Flair, which calculates sentiment using complex deep learning neural networks (Akbik *et al.*, 2019). In order to achieve more representative sentiment estimates and to account for random errors in analysis, two other NLP libraries for Python are also used: The Natural Language Toolkit (NLTK) and TextBlob. To a lesser extent, this study also serves as a demonstration of the utility of these general-purpose sentiment analysis tools in the context of this kind of e-commerce analysis.

1.2 Scope of research

This study covers cross-sectional data on three popular product categories of AliExpress, namely phone screen protectors, phone charging cables, and consumer electronics (a catch-all term that includes various electronics that do not fit other categories). These categories only represent what past studies have referred to as "feature-based goods" (Ghose & Ipeirotis, 2011), or goods whose appeal relies more on functionality over form – a factor which also influences the role of review sentiment, particularly that of subjectivity. The data was gathered in March and April of 2021.

AliExpress was chosen to represent Chinese CBEC, as it is the most popular website in the category (Alexa Internet, 2021) and it is expressly designed to link Chinese sellers to consumers abroad (Xu, 2016). While AliExpress is primarily considered to be a business-to-consumer (B2C) platform, B2B and C2C sales are also supported (Xu, 2016; Liu & Liu, 2017). Because AliExpress has a global reach, and due to limitations in non-English-language sentiment analysis, the analysed reviews were filtered to only include English text.

[1] The number of orders reported by AliExpress per listing only includes recent orders from an unspecified timeframe. The specific duration of this timeframe is not relevant to the results of this study, as it remains constant across all entities observed. Regardless, the duration is likely approximately 6 months, as this is the duration of all other listing data (including reviews/ratings).

Due to the limitations of scope, this study only seeks to examine the relationships between the studied variables, not to predict sales outcomes or establish unambiguous, one-way cause-and-effect connections. From the point of view of the research objectives, this approach is sufficient, as the analysis fits into a broader research context, with the aim of investigating potentially overlooked factors in these relationships. Future study using time-series data would be needed to draw conclusions regarding the causality of these relationships, as well as to account for entity-specific fixed effects.

Despite these limitations in scope, the results of this study will have implications in both expanding on current research and providing valuable findings to a wider business context. Coming to understand the role of user reviews in sales is a fundamentally important exercise for both e-commerce sellers and platform providers alike. The findings of this study allow Western sellers to further understand their increasing competition from the East, and, likewise, cross-border Chinese sellers can come to understand the applicability of previous, Western research in their market. More specific aspects of e-commerce sentiment analysis, such as whether written reviews or numerical ratings are more important, or to what extent subjective statements are associated with increased sales, are also addressed by this study.

1.3 Structure of the research

The remainder of this thesis is structured as follows. Chapter 2 reviews prior literature on the subjects of word-of-mouth in e-commerce, sentiment analysis in sales estimation, and the key aspects of Chinese cross-border e-commerce. Chapter 3 presents the methodological approach for the study, including the specifics of the data collection and processing procedures as well as the construction of the econometric models. Chapter 4 presents the statistical significances, the goodness of fit, and the resultant coefficient estimates of these econometric models. Chapter 5 concludes the paper by further discussing these results, their implications, as well as their limitations and potential for future study.

2 Theoretical background

2.1 Reviews, ratings, and sales

From a theoretical standpoint, marketing and social psychology scholars have long characterised word-of-mouth (WOM) information as a key component of the consumer decision-making process (Brooks, 1957; Kozinets *et al.*, 2010). Traditionally, WOM has been considered a by-product of natural human communication, but with the emergence of online sales platforms and electronic word-of-mouth communication (eWOM), the concept has expanded to include moderated reviews and discussion on the same online sales platforms (Kozinets *et al.*, 2010). Providing users with a native space for user-driven feedback has been shown to drive e-commerce sales by creating an additional information channel and increasing trust (Dellarocas, 2003), and it is no surprise that every mainstream e-commerce platform includes user feedback on its listing pages. Studies have also shown that consumers themselves consider online reviews a significant factor in their purchasing decisions (Lackermair *et al.*, 2013).

Although consumers generally read product evaluations, some debate exists regarding the exact effects of user feedback on consumer behaviour and, subsequently, product sales. In a meta-analysis of the effects of user feedback on sales, Floyd *et al.* (2014) state that, due to a variety of differing conclusions across studies, "a consensus regarding the impact of online product reviews has yet to emerge". Studies such as those by Ghose and Ipeirotis (2006) have found that, while user feedback matters, the average ratings of listings have a less significant effect on sales outcomes than might be expected.

Some of the complexity surrounding the effects of user feedback is a result of inconsistencies in numerical ratings. Factors such as rating inflation, or the gradual increase of user-submitted ratings over time, have devalued the usefulness of ratings in a prospective buyer's decision-making process (Filippas *et al.*, 2018; Garg & Johari, 2020). Due to rating inflation, when the vast majority of listings score above 4½ on a 1 to 5 scale, consumers find it difficult to evaluate the true quality of the product in question (Garg & Johari, 2020; Lee, 2020). Other factors, such as fraudulent reviews, have also played a role – for instance, multiple studies have shown that products and businesses with a star rating between 4.2 and 4.5 tend to outperform those with ratings above 4.5, because users now trust ratings that seem realistic over those that are the highest (Maslowska *et al.*, 2017; Womply Research, 2021). In essence, due to a number of biases in ratings, the average rating of a product is simply not representative of its quality (Hu *et al.*, 2009).

2.2 Sentiment analysis in sales estimation

A potential solution to the issues surrounding rating biases is to analyse customer opinions from textual reviews, instead of using their numerical ratings. Due to the wealth of publicly available textual data, online reviews have been a frequent point of focus in natural language processing. Sentiment analysis approaches have been widely employed in estimating sales and analysing consumer preferences, particularly across software sales (Fu *et al.*, 2013; Maalej & Nabil, 2015; Panichella *et al.*, 2015), the hospitality industry (Duan *et al.*, 2013; Cheng & Jin, 2019), and e-commerce – especially using data from Amazon (Ghose & Ipeirotis, 2006, 2007, 2011; Floyd *et al.*, 2014). The fact that written reviews are very often accompanied by a corresponding numerical rating has also led to researchers using them as convenient datasets to train machine-learning-based sentiment analysis tools (Pang *et al.*, 2002; Pang & Lee, 2008; Bhatt *et al.*, 2015).

Highly cited studies examining the role of analysed sentiment in e-commerce include those by Ghose and Ipeirotis (2006, 2007, 2011), who researched the effect of review subjectivity on demand (using Amazon's Sales Rank as a proxy) and review helpfulness. They found that overall review subjectivity is associated with positive sales outcomes, but for feature-based products, objective reviews were rated as more helpful by other users (Ghose & Ipeirotis, 2011). They suggested that objective reviews mainly verify the validity of the product description (Ghose & Ipeirotis, 2011), potentially reducing perceived risk. More recently, Li *et al.* (2019) approached this subject using a joint sentiment-topic approach on a highly specific dataset of tablet computers, finding that sales amounts can be predicted using either polarity scores or scores for specific product dimensions.

One noteworthy aspect of some of these studies is the potentially false assumption that user-submitted numerical ratings are representative of the theoretical sentiment polarity of the corresponding written reviews. Although Ghose and Ipeirotis acknowledge that "the numeric rating does not capture all the polarity information that appears in the review" (Ghose & Ipeirotis, 2007), they still assume it to be a sufficiently accurate approximation and forgo extracting polarity from the review texts. This assumption is also made by researchers, such as Pang and Lee (2002), who train sentiment analysis models using reviews' ratings as a proxy for their sentiment polarity. However, existing research suggests that this approach is inaccurate, as it goes against the aforementioned research into rating inflation and other similar biases. A large-scale behavioural analysis conducted by Zhang *et al.* (2014) found the assumption to not necessarily be true: users tend to give out higher overall numerical ratings than their reviews would otherwise suggest.

Some researchers have addressed this rating-sentiment-disconnect by treating ratings and text reviews as distinct variables and using alternative data sources to quantify review sentiment polarity. Hu *et al.* (2014) approached this distinction by proposing a sequential decision-making process, in which numerical ratings have an indirect effect on sales by leading potential customers to the product listing. They found that the ratings themselves did not inform the final decision, but that this was instead largely due to the sentiments expressed in the listings' reviews. The findings of Hu *et al.* (2014) suggest that, for user reviews with both plaintext feedback and numerical ratings, the plaintext portion is ultimately the more significant influence on future purchases. On the other hand, Li *et al.* (2019) theorised that, at least across their highly specific data, numerical ratings mediated the overall sentiments expressed in reviews.

In their analysis, Hu *et al.* (2014) used a custom-made sentiment dictionary, which scored sentiment polarity based on common terms like "excellent" or "terrible". Li *et al.* (2019) used an existing lexicon but had to train numerous JST models for their topic identification. These approaches highlight how sentiment analysis tasks can be highly resource-intensive, requiring significant amounts of data and time to train machine learning models. However, basic forms of these sentiment analysis approaches have become increasingly viable as NLP tools have improved, thanks to the emergence of practically applicable general-purpose sentiment analysis tools. This study seeks to utilize the sentiment analysis capabilities of three such (mostly) pre-trained tools, representing a selection of the most popular open-source NLP packages for Python: The Natural Language Toolkit (NLTK), TextBlob, and Flair.

2.3 Chinese cross-border e-commerce

Chinese cross-border e-commerce (CBEC) can be considered to be generally underrepresented in research, in comparison to Western e-commerce and its platforms, such as Amazon. In view of this, previous research has sought to outline some of the distinguishing features of these internationally-oriented Chinese platforms, particularly as compared to their Western and domestic Chinese equivalents. Most notably, the driving success factors of these platforms are their comparatively low prices, wide selections of products, convenience, and overall perceived value (Mou *et al.*, 2017, 2019; Lukicheva & Semenovich, 2019). However, due to bad press and information asymmetries, consumers' purchase decisions on Chinese CBEC platforms are also considerably negatively affected by perceived risk (Mou *et al.*, 2017, 2019).

The lack of trust in Chinese CBEC platforms can result from simple communication inefficiencies due to language and culture barriers (Zhu *et al.*, 2019) or from deeper-ingrained perceived risks with the products or platforms themselves (Mou *et al.*, 2017, 2019). These perceived risks are not always unfounded. For instance, a recent high-profile report by The European Consumer Organisation found that 66 % of products purchased from (largely Chinese) cross-border platforms were unsafe (BEUC, 2021), demonstrating the risks associated with these forms of e-commerce. An analysis of online discussions about AliExpress by Lukicheva and Semenovich (2019) summarized the most commonly perceived risks regarding the platform as resulting from potentially defective or low quality goods, long wait times, and inconsistencies with the declared characteristics of the goods (among other reasons). In short, the role that risk plays in affecting consumers' purchase decisions is considered higher on Chinese CBEC platforms than on Western equivalents.

In order to remain competitive, Chinese CBEC platforms have needed to ensure that their perceived value outweighs any perceived risks. While user reviews have been shown to be an important factor in mitigating perceived risk (Wu *et al.*, 2013), the reviews found on Chinese CBEC platforms are not, themselves, immune from risk, either. Chinese platforms have been found to be particularly prone to seller-reputation-escalation (SRE) practices, known colloquially as "brushing", in which sellers pay for fraudulent reviews (Xu, 2016). These practices potentially exacerbate the aforementioned issue of users distrusting highly-rated listings, which has been shown to result in suboptimal sales outcomes for highly-rated listings (Maslowska *et al.*, 2017). As such, the e-commerce analytics company Profitero has stated that, on Chinese platforms, potential customers have a notable tendency to focus more on review texts than on numerical ratings (Deng, 2016).

Given these differentiating factors, the applicability of the results of Western e-commerce studies on Chinese CBEC platforms is not guaranteed. In particular, due to the increased role of risk and its mitigation, review sentiments can potentially have an even greater relation to sales outcomes than on Western platforms. In addition, due to the increased presence of SRE practices, equating ratings and overall sentiment polarities on Chinese CBEC platforms is potentially particularly erroneous. Other factors of Chinese CBEC platforms, such as the status of rating inflation, are not covered in existing research. As such, this study seeks to address the platform imbalance in e-commerce research, while also accounting for factors that have already been shown to differentiate these contexts.

3 Methodology

3.1 Data collection procedure

To conduct this study, three sets of data from AliExpress product listings were created using publicly available information, including reported order amounts, price options, and review texts. In addition, the approximate ages of the product listings were collected from a separate website, pricearchive.org, that collates AliExpress product listing data. This information was gathered through the use of Python libraries and self-programmed scripts that parse HTML and JavaScript content. The three datasets represent three popular product categories on the website: phone screen protectors (e.g. tempered glass covers), phone charging cables (e.g. USB-C, Micro-USB, and Lightning cables), and consumer electronics (primarily various adapters and audio equipment).

The various types of collected information can be grouped into listing data (listing order amounts, average ratings, and ages), pricing data (various price options) and review data (user-generated written reviews and their corresponding numerical ratings). For each product listing, one instance of all types of listing data, up to 100 price options, and up to 100 reviews were collected. In total, for 10,150 listing data points collected (across all three product categories), 142,494 price options (avg. 14 per listing) and 579,948 written reviews were gathered.

3.1.1 Ethics of data collection

In this data collection context, necessary ethical considerations include those regarding the privacy of the website's users and those regarding the interests of the websites' hosts. Because the results of this study are published in aggregate form, and because no potentially identifying information (e.g. photos, locations, usernames) was collected, the privacy of the website's users was not violated in the data collection process (Kozinets, 2010; Mancosu & Vegetti, 2020). To prevent causing damage to the target websites, the data collection process was conducted in accordance with the best practices and norms of web scraping (Krotov *et al.*, 2020), including by gathering the data in a slow, unobtrusive manner and without violating the respective websites' robots.txt guidelines. It is also worth noting that the gathered data is already publicly available and that all of the collected data was promptly deleted after analysis and neither published nor otherwise used to create a database.

3.2 Data preparation and variables

Once collected, the three sets of raw data were processed into variables for regression. The selection of variables for regression was a product of the data available and previous work set by Chevalier and Mayzlin (2006) as well as Ghose and Ipeirotis (2006, 2007, 2011). To analyse the role of review sentiment, the variables *polarity* and *subjectivity* were calculated from the raw review texts. In addition, to align with previous works, a unified, representative *price* variable (for listings with multiple price options) was determined. Other control variables, such as those accounting for pricing-related search result advantages, were also calculated. Rows with null values (primarily those without age estimates on Pricearchive) were filtered out of the analysis, resulting in a total of 8,319 listings across all three product categories. Non-English-language reviews were filtered out using langdetect, a Python port of a Java-based language-detection library (Danilak, 2020), resulting in a total of 451,375 reviews to be analysed across all listings.

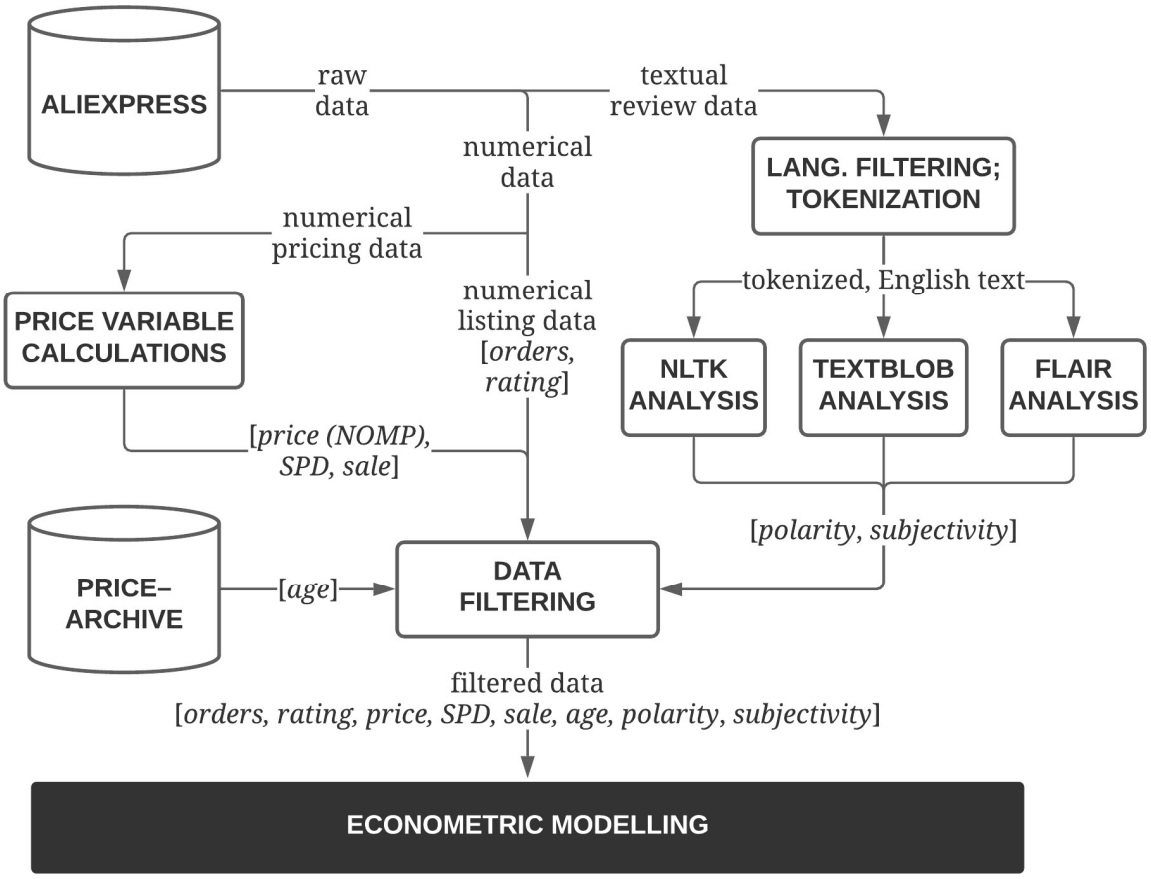


Figure 1. System design of the data preparation process

3.2.1 Listing variables

First, for each product listing, a handful of directly available "listing" variables were collected. These included the variable *orders* for the number of orders (used as the dependent variable in empirical analysis) and the variable *rating* for the average rating (on a scale of one to five stars) of the listing. In addition, the variable for the estimated age (*age*) of the listing was retrieved. Because AliExpress does not directly report the age of the product listing, the *age* values were collected from Pricearchive.org, a third-party website that collates and compares historical pricing information on AliExpress listings (Pricearchive.org, 2020). These ages are considered estimates, as the values represent the earliest date from which Pricearchive has data on a given listing; in practice, the listings are likely slightly older, but this difference can be expected to be minimal and roughly constant across all examined listings. Listings that did not have age estimates on Pricearchive were filtered out of the analysis (approx. 10 % of all listings).

It is important to note that AliExpress does not report a listing's order amount throughout its history – instead, it only reports the order amount across an unspecified duration. Based on the fact that all other time-relevant information, including user reviews and seller ratings, is limited to the past 6 months, it could be assumed that the order amounts are from the past 6 months, as well. This means that a listing's age is likely to have a smaller impact on the order amounts than one might expect – especially because the vast majority of listings (approx. 88 %) were older than 6 months. Regardless, due to indirect factors (such as third-party recommendations for and links to specific listings), this variable may have some impact and was still included in the analysis.

3.2.2 Pricing variables

The pricing information of each listing was gathered with up to 100 price options representing the various colour/size/etc. combinations available for the product in question. Because AliExpress does not present information on how many of the listing's orders were placed at whichever price option, some assumptions had to be made about the most representative price for each listing. If the various prices for a single listing did not differ considerably, this representative price could have been calculated as a simple arithmetic mean of the price options; however, a cursory examination of product listings revealed that price options can vary significantly, and care needed to be taken to select a price that would meaningfully represent the listing as a whole.

Most notably, a significant portion of the listings included one or two price options that were considerably lower than the other options and did not entirely match the title of the listing. A typical example from consumer electronics is a listing for a roughly 78 – 112 USD battery management system that also includes an option to buy a 4 USD Bluetooth module (see Figure 2). We can reasonably expect most consumers who click open the listing to be in search of a battery management system, so the majority of them will not be purchasing the Bluetooth module (at least not exclusively). This makes the 4 USD price option an outlier, meaning that an arithmetic mean of all price options would skew lower than an ideal, representative price.

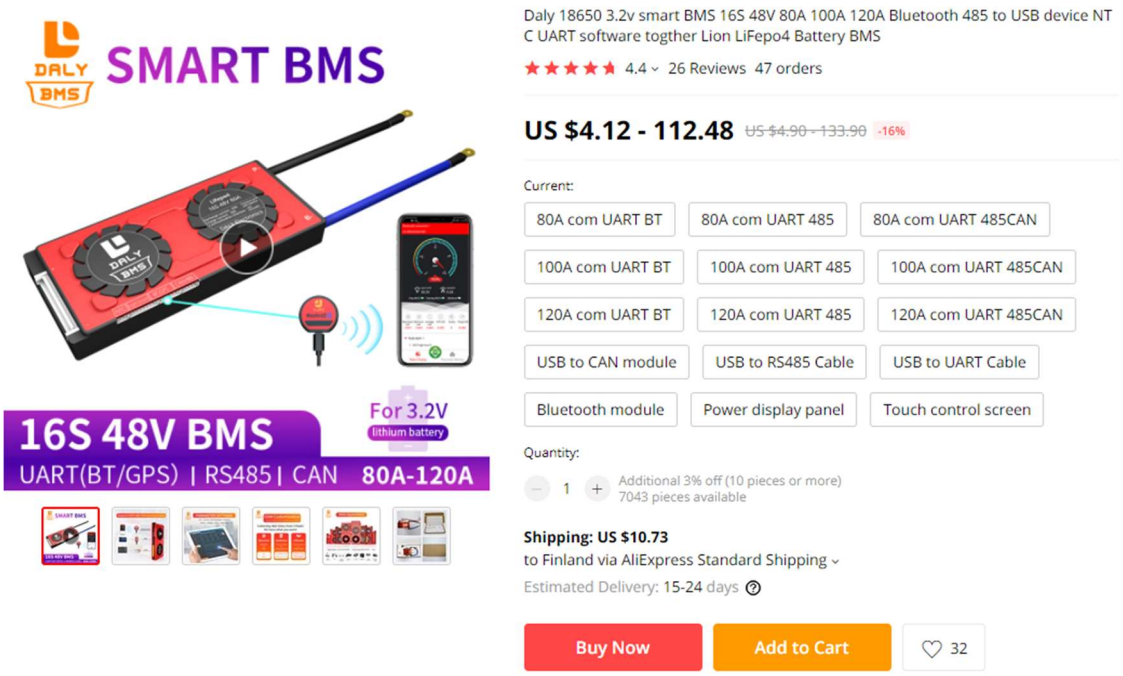


Figure 2. Example listing with outlier price options (AliExpress, n.d.)

Using a simple median of the price options would avoid the influence of these low outliers, but, as stated by Lukicheva and Semenovich (2019), a primary draw of AliExpress is its price competitiveness. Based on this, we can assume that many of the platform's customers are price-sensitive and may be drawn towards the lowest acceptable price option, i.e. the cheapest option that still meets their expectations. In the example of the battery management system, this lowest acceptable price option would be the one around 78 USD – neither the minimum option of 4 USD (which does not meet the usual customer's expectations) nor the median option of approximately 97 USD (which is a higher price than a highly price-sensitive customer would select).

To adjust for this issue, for each listing, a subset of price options was created that excluded any prices that were defined as statistical outliers according to the Median Absolute Deviation (MAD) method, as recommended by Leys *et al.* (2013). This meant that a representative, non-outlier minimum price (to be used as the *price* variable in regression analysis) could be identified. Approximately 19 % of all listings had a non-outlier minimum price that differed from the minimum of all listed price options. The formula used for calculating MAD is as follows (Huber, 1981, p. 107):

$$MAD = b \times \text{med}\{|x_i - \text{med}\{x_i\}|\}$$

where $b = 2.5$ (standard multiplier suggested by Leys *et al.* (2013)) and x_i represents each of the price options. Using MAD, each listing's non-outlier minimum price ($NOMP$) was calculated as follows:

$$NOMP = \min\{x_i\}, \text{ such that } \min\{x_i\} \geq \text{med}\{x_i\} - MAD$$

However, (low) outlier price options still influence the visibility of the listing in search results. In the example of the battery management system, the price of the listing appears as 4 – 112 USD in search results, so while most users will not have spent 4 USD on the product, it still appears as 4 USD when sorting by lowest price. To account for this factor, a separate variable, the Search Price Distortion (SPD) was created. The SPD describes the competitive advantage that a listing achieves (compared to other comparable listings) by including a low outlier price. It is a simple percentage difference between the minimum price and the non-outlier minimum price, according to the following formula:

$$SPD = 1 - (\min\{x_i\} \div NOMP)$$

In the battery management system example above, given that $NOMP$ is roughly 78 USD and $\min\{x_i\}$ is 4 USD, the SPD is roughly 95 % – a relatively extreme case. For the 81 % of listings without outlier prices, SPD is zero. Considering the findings of Hu *et al.* (2014), who define a distinction between factors that lead a customer to a product listing and factors that influence their final purchasing decision, we can consider SPD to represent the former category and $NOMP$ to represent the latter (and the former, to some degree).

Finally, to factor for the effects of any reported discounts, the variable *sale* was calculated. It is the percentage difference between the actual, discounted price and the given original price. These sales percentages are constant across all of a listing's price options.

$$sale = 1 - (price_{actual} \div price_{original})$$

3.2.3 Review variables

For each listing, a maximum of 100 reviews was collected, based on the order in which AliExpress presented them by default. Filtered for exclusively English-language reviews, this gave a total of 451,375 reviews (across all listings) for sentiment analysis. In order to alleviate random errors in machine-learning-based sentiment analysis, three different tools were used in parallel. These were three of the most notable open-source NLP libraries for Python: The Natural Language Toolkit (NLTK) (Loper & Bird, 2002), TextBlob (Loria, 2020), and Flair (Akbik *et al.*, 2019). The resulting polarity and subjectivity values from these tools' calculations were averaged for more representative overall scores.

First, for each sentence of a given review, a sentiment polarity score was calculated. All three libraries include a pre-trained sentiment polarity analysis algorithm, with both NLTK's Valence Aware Dictionary for sEntiment Reasoning (VADER) and TextBlob using simple rule-based models (Hutto & Gilbert, 2014; Loria, 2020) and Flair using a more complex character-level deep learning neural network (Akbik *et al.*, 2019). All of these algorithms rate sentiment polarity on a scale of -1 to +1. The resultant sentence-level polarity scores were averaged to achieve a review-level polarity score, and finally, these review-level polarity scores were averaged to achieve a listing-level overall polarity score, assigned to the listing's *polarity* variable.

Second, in addition to sentiment polarity, the three tools were used to quantify the overall subjectivity of each sentence of each review. TextBlob comes with a pre-trained subjectivity classifier algorithm, but for NLTK and Flair, new subjectivity classifiers had to be trained. Both NLTK and Flair include source code links to subjectivity corpora – in both cases, these were Pang and Lee's (2004) set of 5,000 subjective and objective sentences pulled from movie reviews and plot descriptions. For NLTK, a comparatively simple naïve Bayes classifier was used, while for Flair, a more complex Transformer model (distilBERT) was used, as these were the default options outlined in the packages' documentations. Once trained, the NLTK classifier would rate subjectivity binarily as either "obj" or "subj" (translated to 0 and 1, respectively), while both Flair and TextBlob would rate subjectivity on a scale of 0 to 1 (with 1 representing 100 % confidence that the given statement is subjective). As with the polarity scores, the sentence-level subjectivity scores were first averaged to assign a subjectivity score to each review, and then these review-level subjectivity scores were averaged to achieve a listing-level overall subjectivity score, assigned to the listing's *subjectivity* variable.

Prior to calculating polarity and subjectivity, however, the review texts themselves had to be broken down into sentence tokens and, for the custom-trained NLTK subjectivity classifier, word tokens. This way, each sentence of a given review is analysed individually, with the NLTK subjectivity classifier interpreting the sentences as a "list of strings" data structure. Other common NLP pre-processing techniques, such as stopword removal and lemmatization, were not used, as the three NLP packages are sufficiently context-aware to account for these factors in sentences.

The polarity and subjectivity were analysed on a sentence-by-sentence basis, because NLTK's VADER (as well as the custom-trained NLTK and Flair subjectivity classifiers) was trained using sentence-level text snippets (Hutto & Gilbert, 2014), resulting in potentially inaccurate results when analysing multi-sentence text snippets. Because the sentence-level scores were averaged for each review, each review's sentiment is equivalently weighted in the overall *polarity* and *subjectivity* variables for the listing, regardless of the number of sentences in the review. In essence, the *polarity* and *subjectivity* variables can be considered to represent the probable polarity and subjectivity of any randomly selected user review for a given listing.

By charting the sentiment analysis results for each of the three NLP packages, the distributions of the reviews can be observed. As can be seen from Appendix 1, both NLTK's VADER and TextBlob rate polarity quite modestly, whereas Flair is considerably more confident in its results. This means that the resulting averages of the three values will be spread wider than when only using NLTK's VADER and TextBlob. By contrast, both of the custom-trained subjectivity models (NLTK and Flair) rate the majority of reviews as entirely subjective (although for NLTK this figure is more extreme, due to the binary nature of its classification), whereas TextBlob leans towards them being mostly objective. There are no notable differences in these trends between the three product categories.

By grouping the reviews by their corresponding user-submitted numerical ratings, the prevalence of rating inflation can also be seen. As can be seen from the review heatmap in Appendix 3, the vast majority of all reviews (approx. 88 %) had a numerical rating of 5 stars. This disparity is so severe that even the 5-star reviews with negative sentiment polarities outnumbered all 1 to 4-star reviews.

3.3 Empirical econometric modelling

$$\begin{aligned}\ln(orders)_i = \alpha + \\ \beta_1 \ln(price)_i + \\ \beta_2 \ln(age)_i + \\ \beta_3\, rating_i + \\ \beta_4\, sale_i + \\ \beta_5\, SPD_i + \\ \beta_6\, polarity_i + \\ \beta_7\, subjectivity_i + \\ \varepsilon_i\end{aligned}$$

For each product category, the cross-sectional econometric model above was estimated using ordinary least squares (OLS) regression. In the model, α is the model's intercept term, β_1 to β_8 are the coefficients of the regressors (explanatory variables), and ε_i is the error term. The subscript i represents a given instance of the different entities (listings) observed. Below, Tables 1 and 2 display descriptive information about these variables.

Table 1. Definitions of variables

Variable	Definition	Range min	Range max
$\ln(orders)$	Natural logarithm of the number of orders	n/a	n/a
$\ln(price)$	Natural logarithm of the non-outlier minimum price (NOMP)	n/a	n/a
$\ln(age)$	Natural logarithm of the number of days the listing has existed	n/a	n/a
$rating$	Average numerical rating of the listing	1	5
$sale$	Reported discount percentage of the listing	0 %	100 %
SPD	Percentage difference between minimum price and NOMP	0 %	100 %
$polarity$	Average sentiment polarity (-1 to +1) of the user reviews	-1	+1
$subjectivity$	Average subjectivity (0 to 1) of the user reviews	0	+1

Table 2. Summary statistics

	Screen protectors		Phone charging cables		Consumer electronics	
Variable	Mean	St. Dev.	Mean	St. Dev.	Mean	St. Dev.
$orders$	2925.81	6616.77	1462.93	7035.06	177.68	445.18
$price$	1.18	1.33	2.32	2.79	15.33	48.00
age	394.31	253.19	441.20	309.41	470.25	288.31
$rating$	4.70	0.12	4.79	0.14	4.81	0.17
$sale$	54.83 %	34.57 %	43.73 %	29.95 %	22.36 %	17.85 %
SPD	9.81 %	16.84 %	10.79 %	22.97 %	2.87 %	12.02 %
$polarity$	0.28	0.11	0.34	0.13	0.33	0.18
$subjectivity$	0.66	0.04	0.68	0.05	0.68	0.8

Notes: While the variables $orders$, $price$, and age are transformed into their natural logarithms for regression, for easier conceptualization, they are presented here as-is.

4 Results

Table 3 presents the results of OLS regression for each of the three product categories. For correlation matrices and variable-specific VIF values, see Appendix 2.

Table 3. OLS regression estimates

Coefficient / variable		Screen protectors		Phone charging cables		Consumer electronics	
		Estimate	Std. Err.	Estimate	Std. Err.	Estimate	Std. Err.
α	$Intercept$	9.1436 ***	0.9773	8.7174 ***	0.7650	4.5984 ***	0.5833
β_1	$\ln(price)$	−0.0770 **	0.0251	−0.2106 ***	0.0181	−0.0734 ***	0.0112
β_2	$\ln(age)$	−0.0903 **	0.0327	0.1337 ***	0.0243	0.0272	0.0267
β_3	$rating$	−0.0724	0.1802	−1.0553 ***	0.1535	−0.1253	0.1140
β_4	$sale$	−0.4287 *	0.1753	0.9197 ***	0.1336	1.4468 ***	0.1006
β_5	SPD	1.0008 ***	0.1363	0.3338 ***	0.0846	0.5490 ***	0.1409
β_6	$polarity$	3.5641 ***	0.2131	2.2061 ***	0.1975	0.4554 ***	0.1187
β_7	$subjectivity$	−2.9110 ***	0.4993	0.3304	0.4072	−0.0472	0.2415
Observations			2,295		3,229		2,795
R^2			0.1605		0.3650		0.1225
Adjusted R^2			0.1579		0.3637		0.1203
F-statistic			438.63		1856.4		390.36
Maximum VIF			10.897		5.1519		1.7164

Notes: The dependent variable is $\ln(orders)$. Across coefficient estimates, ***, **, and * denote statistical significances at 0.1 %, 1 %, and 5 %, respectively.

The F-test indicates that the overall OLS results for all three product categories are significant at the $p < 0.0001$ significance level. The parameters which are consistently significant ($p < 0.01$) across all three groups are α, β_1, β_5, and β_6. The remaining parameters, β_2, β_3, β_4, and β_7, are inconsistent in their estimates and/or significances, which could be due to differences between product groupings, limitations in data, or other inaccuracies in model estimation.

The coefficients of determination vary noticeably between the three groups, demonstrating that the degree to which changes in the dependent variable are reflected in changes in the independent variables is not consistent across groupings. In essence, this variation aligns with the specificity of the product category in question; consumer electronics is the broadest grouping, so the number of potential unaccounted factors affecting order amounts is greater. By contrast, phone charging cables (the specifications of which do not vary much) are most similar, and a greater degree of the change in orders is associated with these listing-specific independent variables. These differences across product categories are discussed further in Section 5.

The maximum variance inflation factors[2] demonstrate a degree of multicollinearity in the first two models, with $\ln(price)$ and $sale$ having VIF values greater than 2.5 among screen protectors and phone charging cables, indicating "considerable collinearity" (Johnston *et al.*, 2018). These VIF values can be found in Appendix 2d. To correct for multicollinearity, an additional set of regression models was estimated, with β_4 $sale$ excluded from the analysis:

Table 4. Multicollinearity-corrected OLS regression estimates with β_4 $sale$ excluded

Coefficient / variable		Screen protectors		Phone charging cables		Consumer electronics	
		Estimate	Std. Err.	Estimate	Std. Err.	Estimate	Std. Err.
α	$Intercept$	9.5373 ***	0.9652	8.6877 ***	0.7706	5.8683 ***	0.5976
β_1	$\ln(price)$	−0.0198 *	0.0091	−0.3212 ***	0.0084	−0.1294 ***	0.0109
β_2	$\ln(age)$	−0.0922 **	0.0328	0.1158 ***	0.0244	−0.0107	0.0275
β_3	$rating$	−0.1569	0.1771	−0.9906 ***	0.1543	−0.2834 *	0.1176
β_5	SPD	0.9782 ***	0.1362	0.4642 ***	0.0831	0.7382 ***	0.1454
β_6	$polarity$	3.5715 ***	0.2133	2.2049 ***	0.1990	0.4839 ***	0.1230
β_7	$subjectivity$	−3.1305 ***	0.4918	0.5895	0.4084	0.1165	0.2500
Observations			2,295		3,229		2,795
R^2			0.1583		0.3557		0.0576
Adjusted R^2			0.1561		0.3545		0.0556
F-statistic			431.52		1782.7		170.85
Maximum VIF			1.4631		1.8704		1.7159

Notes: The dependent variable is $\ln(orders)$. Across coefficient estimates, ***, **, and * denote statistical significances at 0.1 %, 1 %, and 5 %, respectively.

Naturally, as this multicollinearity was not visible in consumer electronics, its coefficient of determination and F-statistic are decreased considerably from the previous model. For the other two product categories, these statistics do not show a large change. As would be expected, of the leftover coefficients, the largest change occurred in the collinear variable $price$, while the others remained relatively consistent.

Notably, neither significant collinearity nor excessive correlation is observed between the variables $rating$ and $polarity$, demonstrating that these two variables differ noticeably. This indicates that a given product listing's average rating is not entirely linearly representative of the overall polarity of its written reviews, and vice versa.

[2] The variance inflation factor (VIF) is the reciprocal of the tolerance $(1 - R^2)$ of a model in which a given explanatory variable is regressed on all other explanatory variables included in the original analysis (Johnston *et al.*, 2018). It is calculated for all explanatory variables, and the maximum of these VIFs serves as an indicator of multicollinearity in the original model.

5 Discussion and conclusions

Although the relationships between e-commerce user reviews and product sales have been extensively covered in existing research, prior studies have largely focused on data from a limited selection of platforms. In addition, many of these studies have drawn equivalences between user-submitted ratings and the polarities of corresponding textual reviews, conflating the two under a singular "valence" term. Using data from a context relatively underrepresented in research, this study demonstrates that some of the previous studies' findings can be generalized to the Chinese cross-border e-commerce context, while some others cannot. The disconnect between numerical ratings and written reviews is also visible in these results, reinforcing the notion that users tend to give higher numerical ratings than their written opinion would otherwise imply (Zhang *et al.*, 2014).

As mentioned in Section 4, an important distinction between these product categories is the breadth of the types of products they represent. Of the three groupings, the broadest is consumer electronics, which is a "catch-all" categorisation that includes various niche products. This means that the factors affecting a potential customer's purchase decision are numerous, and accounting for them in a cross-sectional regression analysis is not practical. By contrast, the most specific grouping is phone charging cables. Due to the standardization of phone charging inputs, there are commonly only three types of sockets (USB-C, micro-USB, and Lightning), and most consumers looking to purchase a charging cable can have their demands met by almost any listing. This increases the relative importance of factors like pricing and user reviews. In between these two categories is phone screen protectors, the listings of which are similar to each other but dependent on the model of the phone in question, so listings representing protectors for more popular phones will have a considerable advantage in order amounts. Understanding this distinction is key to understanding the generalizability of the results as a whole.

Prior to interpreting the specifics of the results, the role of algorithmic bias in the retrieval of review texts should be addressed. Because a maximum of 100 written reviews was collected per listing, listings with more reviews than this may be subject to bias, as the sampling is not necessarily random. AliExpress sorts user reviews algorithmically, possibly favouring some kinds of reviews over others. Regardless, a quick comparison of listings with more than 100 reviews and those with fewer than 100 shows that the correlation between sentiment polarities (calculated from up to 100 reviews) and average ratings (from all reviews, as reported by AliExpress) is not significantly different, suggesting that AliExpress does not bias its review sorting in a way that affects these results.

5.1 Implications to research

First, examining the analysed sentiment of reviews in Tables 3 and 4, the positive relationship between the overall sentiment polarity and the number of orders for a given listing is found to be relatively consistent across all three product groupings. This relationship is both intuitively reasonable and aligned with prior research, as positive reviews have repeatedly been shown to encourage further sales (Floyd *et al.*, 2014). In addition, both review positivity and sales outcomes may be jointly associated with uncontrolled variables, such as the overall appeal of the product in question. In any case, it can be determined that review valence has a similarly positive relationship with order amounts in the context of Chinese CBEC as it does on Western platforms. In general, it can be understood that "good" products are simultaneously more sought after and more likely to receive positive reviews, regardless of the platform in question.

The role of subjectivity in reviews is less consistent. Between the OLS estimates of the three product groupings, subjectivity was only statistically significant in one: phone screen protectors. In this category, overall subjectivity had a negative relationship with order amounts, suggesting that listings with more objective, fact-based reviews are also in higher demand. While this goes against the general findings of Ghose and Ipeirotis (2006, 2007, 2011), they do note that, for feature-based goods (such as the products studied here), users tend to find objective reviews more helpful. This could mean that, at least for feature-based products, user reviews serve a role in mitigating risk by confirming the product description. Seeing that perceived risk is an exceptionally important factor in consumers' decision-making processes on Chinese platforms (Mou *et al.*, 2017, 2019), this risk mitigation effect of review objectivity may outweigh the benefits of subjectivity observed on Western platforms.

Notably, the average rating of a product listing is not consistently statistically significant, suggesting that potential customers pay more attention to written reviews than numerical ratings. This is in accordance with the findings of Hu *et al.* (2014). Where it is significant, the average rating actually has a negative relation to order amounts – while this may seem unintuitive, it aligns with previous research demonstrating that unrealistically high ratings discourage purchases (Maslowska *et al.*, 2017). Examining the connection between rating and polarity on a review level, it is worth noting that the polarity of a given review and its corresponding numerical rating is somewhat correlated at an approximate correlation coefficient of 0.47, but, at the same time, an overwhelming majority of reviews have a rating of 4 or 5 stars – regardless of polarity. Across all reviews

analysed, there were more 5-star ratings with a negative review sentiment than there were 1 to 4-star ratings altogether (see Appendix 3). This seems to suggest that rating inflation is a significant factor in the Chinese CBEC context – perhaps even more so than on Western platforms.

In terms of control variables, the price of a listing (*price*) is consistently negatively related to the number of orders, aligning with the notion that AliExpress shoppers are price-sensitive (Lukicheva & Semenovich, 2019). In addition, the Search Price Distortion variable (*SPD*), which controls for products with outlier minimum prices, is positively related to the number of orders, suggesting that the search ranking improvements associated with outlier prices are beneficially related to increased sales. In the model for consumer electronics, where it did not cause multicollinearity, the discount percentage (*sale*) was positively related to the number of orders, which makes intuitive sense, as discounts make the listing's price seem more appealing. In general, it can be stated that the control variables behave as would be expected based on prior research.

5.2 Implications to practice

From a practical perspective, the results of this study have the effect of demonstrating that previous research into sentiment analysis in Western e-commerce is, broadly speaking, applicable to the Chinese CBEC context. This means that companies currently engaged in or looking to transition to Chinese CBEC can utilize much of the existing research, without fear that the markets are too different. Some of the more nuanced differences in the results of this study, such as the seemingly different role of subjectivity in reviews, should also be accounted for.

Overall, the understanding that review sentiment can be empirically demonstrated to be significantly related to order amounts is an important starting point for any developments involving user reviews. This study confirms that negative review sentiments can harm sales and that it is crucial for e-commerce businesses to monitor these sentiments. In addition, the fact that these general-purpose sentiment analysis tools provided meaningful results is useful: platforms and sellers can use the same tools in their own sentiment analysis tasks and expect to get results, at least on a large scale. Among other purposes, platforms and sellers can use these tools to quickly identify and address unsatisfied customers, especially those that leave high numerical ratings with a negative review text. On a larger scale, companies can also use these calculated sentiment metrics as performance indicators to compare listings and make strategic decisions.

5.3 Limitations and future research

While the results of this study are based on large amounts of empirical data, some fundamental limitations exist in their applicability. In order to build on the results of this study and to apply its results to a broader context, an understanding of these limitations is necessary.

One important consideration regarding these results is that the polarity and subjectivity scores calculated using sentiment analysis tools are not necessarily objective measurements – they are the "best guesses" of complex algorithms trained on large datasets. While these tools can sometimes outperform humans (Hutto & Gilbert, 2014), their performance is context-dependent, and their accuracy/precision in this study cannot be estimated without manual review. As such, the results of this study are best understood and utilized in the context of general-purpose sentiment analysis, with the caveat that discrepancies between NLP-assigned sentiments and theoretical "true" sentiments can exist. Regardless, this study's use of three different NLP libraries can be expected to have reduced random error, improving reproducibility with different data.

Another limitation regarding the analysed sentiment scores is the relative simplicity of the chosen variables. Aspects such as the overall length and readability of the reviews were not included, meaning that each review was treated with equal weight. This could cause some distortion, as potential customers are likely to assign greater weight to reviews with more substance. In addition, the aforementioned algorithmic bias could affect this, as reviews displayed first by AliExpress will have a greater impact on purchase decisions than those found further into the feedback section. The role of photos in reviews was also left out of this study, even though review photos could be expected to have a large impact on the mitigation of perceived risk.

Overall, the generalizability of these results is limited by the relatively specific groupings of products analysed. Although limiting the research to specific categories improves the significances of the analysed variables, additional research would have to be conducted to confirm that their estimates hold true for other product categories, as well. In particular, the products analysed in this study were largely feature-based, and factors such as review subjectivity will likely have differing effects on more subjectively-differentiated product categories, such as clothing or decorations. On an even broader note, the generalizability of these results is possibly affected by the limitation of a single e-commerce data source; for future studies, additional CBEC platforms should also be investigated.

Finally, and perhaps most significantly, the results of this study are limited in that they do not seek to establish any direct causal links between the analysed variables. Due to the limited scope of this study, data could not be gathered over significant time spans, and thus this analysis is purely cross-sectional. This means that the results cannot be used to predict future sales outcomes or account for potential simultaneity in the model, resulting in some concerns regarding endogeneity. To account for this, some of the variables from other studies which could possibly contribute to simultaneity in a cross-sectional model, such as review counts[3], were deliberately left out of this study. The generalizability of the results may also be limited due to effects such as seasonality. Given additional time resources, further study could be conducted by following the same product listings over the course of multiple years, allowing for more comprehensive predictive time series analysis and a better understanding of trends and seasonal effects. Several time observations would also allow for a more robust entity-demeaned fixed effects model, enabling control over heterogeneity across products, as well.

Despite these limitations, the results of this study can be considered to give valid and significant answers to the outlined research questions. By and large, these limitations invite future study and deeper analysis, for an even better understanding of the ever-growing field of e-commerce.

[3] While the number of reviews for a given product listing likely affects the number of orders, the reverse is true, as well. In other words, people are more likely to buy products with many reviews, but new orders also lead to more reviews. This means that, for cross-sectional data, including the number of reviews as a variable is likely to contribute to significant simultaneity.

References

Akbik, A. *et al.* 2019. 'FLAIR : An Easy-to-Use Framework for State-of-the-Art NLP'. in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. Minneapolis, MN, USA: Association for Computational Linguistics. P. 54–59. (DOI: 10.18653/v1/N19-4010).

Alexa Internet. 2021. *The top 500 sites on the web*. *alexa.com*. [Last accessed 11 Mar. 2021] Available at: https://www.alexa.com/topsites.

AliExpress. n.d. *Daly 18650 3.2v smart BMS 16S 48V 80A 100A 120A Bluetooth 485 to USB device NTC UART software togther Lion LiFepo4 Battery BMS*. *aliexpress.com*. [Last accessed 31 Mar. 2021] Available at: https://www.aliexpress.com/item/4001321708781.html.

BEUC. 2021. *Is it safe to shop on online marketplaces ?* Available at: https://www.beuc.eu/publications/beuc-x-2021-004_is_it_safe_to_shop_on_online_marketplaces.pdf.

Bhatt, A. *et al.* 2015. 'Amazon Review Classification and Sentiment Analysis'. *International Journal of Computer Science and Information Technologies*. 6(6). P. 5107–5110

Brooks, R. C. 1957. '"Word-of-Mouth" Advertising in Selling New Products'. *Journal of Marketing*. 22(2). P. 154–161. ISSN 0022-2429. (DOI: 10.1177/002224295702200205).

Cheng, M. & Jin, X. 2019. 'What do Airbnb users care about? An analysis of online review comments'. *International Journal of Hospitality Management*. 76(May 2018). P. 58–70. ISSN 02784319. (DOI: 10.1016/j.ijhm.2018.04.004).

Chevalier, J. A. & Mayzlin, D. 2006. 'The effect of word of mouth on sales: Online book reviews'. *Journal of Marketing Research*. 43(3). P. 345–354. ISSN 00222437. (DOI: 10.1509/jmkr.43.3.345).

Danilak, M. 2020. *langdetect 1.0.8*. *pypi.org*. [Last accessed 1 Mar. 2021] Available at: https://pypi.org/project/langdetect/.

Dellarocas, C. N. 2003. 'The Digitization of Word-of-Mouth: Promise and Challenges of Online Feedback Mechanisms'. *SSRN Electronic Journal*. ISSN 1556-5068. (DOI: 10.2139/ssrn.393042).

Deng, Y. 2016. *Singles' Day: How Brands Can Win in China's eCommerce Market. profitero.com*. [Last accessed 11 Mar. 2021] Available at: https://www.profitero.com/blog/2016/11/singles-day-how-brands-can-win-in-chinas-ecommerce-market.

Duan, W. *et al.* 2013. 'Mining online user-generated content: Using sentiment analysis technique to study hotel service quality'. *Proceedings of the Annual Hawaii International Conference on System Sciences*. P. 3119–3128. ISSN 15301605. (DOI: 10.1109/HICSS.2013.400).

Filippas, A., Horton, J. J. & Golden, J. 2018. 'Reputation inflation'. *ACM EC 2018 - Proceedings of the 2018 ACM Conference on Economics and Computation*. P. 483–484. ISBN 9781450358293. (DOI: 10.1145/3219166.3219222).

Floyd, K. *et al.* 2014. 'How online product reviews affect retail sales: A meta-analysis'. *Journal of Retailing*. 90(2). P. 217–232. ISSN 00224359. (DOI: 10.1016/j.jretai.2014.04.004).

Fu, B. *et al.* 2013. 'Why people hate your App - Making sense of user feedback in a mobile app store'. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Part F1288. P. 1276–1284. ISBN 9781450321747. (DOI: 10.1145/2487575.2488202).

Garg, N. & Johari, R. 2020. 'Designing Informative Rating Systems: Evidence from an Online Labor Market'. in *Proceedings of the 21st ACM Conference on Economics and Computation*. New York, NY, USA: ACM. ISSN 23318422. (DOI: 10.1145/3391403.3399455).

Ghose, A. & Ipeirotis, P. G. 2006. 'Designing ranking systems for consumer reviews: The impact of review subjectivity on product sales and review quality'. *16th Workshop on Information Technologies and Systems, WITS 2006*. (June 2014). P. 217–222

Ghose, A. & Ipeirotis, P. G. 2007. 'Designing novel review ranking systems'. *ICEC '07: Proceedings of the ninth international conference on Electronic commerce*. P. 303–310. ISBN 9781595937001. (DOI: 10.1145/1282100.1282158).

Ghose, A. & Ipeirotis, P. G. 2011. 'Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics'. *IEEE Transactions on Knowledge and Data Engineering*. 23(10). P. 1498–1512. ISSN 10414347. (DOI: 10.1109/TKDE.2010.188).

Hu, N., Koh, N. S. & Reddy, S. K. 2014. 'Ratings lead you to the product, reviews help you clinch it? the mediating role of online review sentiments on product sales'. *Decision Support Systems*. 57(1). P. 42–53. ISSN 01679236. (DOI: 10.1016/j.dss.2013.07.009).

Hu, N., Zhang, J. & Pavlou, P. A. 2009. 'Overcoming the J-shaped distribution of product reviews'. *Communications of the ACM*. 52(10). P. 144–147. ISSN 00010782. (DOI: 10.1145/1562764.1562800).

Huber, P. J. 1981. *Robust Statistics*. Hoboken, NJ, USA: John Wiley & Sons, Inc. (Wiley Series in Probability and Statistics). ISBN 0-471-41805-6. (DOI: 10.1002/0471725250).

Hutto, C. J. & Gilbert, E. 2014. 'VADER: A parsimonious rule-based model for sentiment analysis of social media text'. *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014*. June 2014. P. 216–225. ISBN 9781577356578.

Johnston, R., Jones, K. & Manley, D. 2018. 'Confounding and collinearity in regression analysis: a cautionary tale and an alternative procedure, illustrated by studies of British voting behaviour'. *Quality & Quantity*. 52(4). P. 1957–1976. ISSN 0033-5177. (DOI: 10.1007/s11135-017-0584-6).

Kozinets, R. V. 2010. *Netnography: Doing Ethnographic Research Online*. SAGE Publications. ISBN 1848606451.

Kozinets, R. V *et al.* 2010. 'Networked Narratives: Understanding Word-of-Mouth Marketing in Online Communities'. *Journal of Marketing*. 74(2). P. 71–89. ISSN 0022-2429. (DOI: 10.1509/jm.74.2.71).

Krotov, V., Johnson, L. R. & Silva, L. 2020. 'Tutorial: Legality and Ethics of Web Scraping'. *Communications of the Association for Information Systems*. 47. P. 555–581. (DOI: 10.17705/1CAIS.04724).

Lackermair, G., Kailer, D. & Kanmaz, K. 2013. 'Importance of Online Product Reviews from a Consumer's Perspective'. *Advances in Economics and Business*. 1(1). P. 1–5. (DOI: 10.13189/aeb.2013.010101).

Lee, G. J. K. 2020. *Comparing Numerical Ratings and Plain-Text Feedback from Online Reputation System: Evidence from Sentiment Analysis of Airbnb reviews in London*. *SSRN Electronic Journal*. Columbia University Graduate School of Arts and Sciences. (DOI: 10.2139/ssrn.3611064).

Leys, C. *et al.* 2013. 'Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median'. *Journal of Experimental Social Psychology*. 49(4). P. 764–766. ISSN 00221031. (DOI: 10.1016/j.jesp.2013.03.013).

Li, X., Wu, C. & Mai, F. 2019. 'The effect of online reviews on product sales: A joint sentiment-topic analysis'. *Information and Management*. 56(2). P. 172–184. ISSN 03787206. (DOI: 10.1016/j.im.2018.04.007).

Liu, X. & Liu, R. 2017. 'The Comparison and Analysis of China Cross-border E-commerce Business Model'. in *Proceedings of the 2017 International Conference on Information Technology and Intelligent Manufacturing (ITIM 2017)*. Atlantis Press. P. 85–88. (DOI: 10.2991/itim-17.2017.22).

Loper, E. & Bird, S. 2002. 'NLTK: The Natural Language Toolkit'. *Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*. P. 62–69. Available at: https://arxiv.org/abs/cs/0205028.

Loria, S. 2020. *TextBlob: Simplified Text Processing – TextBlob 0.16.0 documentation*. *readthedocs.io*. [Last accessed 6 Apr. 2021] Available at: https://textblob.readthedocs.io/en/dev/index.html.

Lukicheva, T. & Semenovich, N. 2019. 'Big Data as a success factor of AliExpress in the Russian market: advantages and opportunities as seen by the eyes of consumers'. in *Third International Economic Symposium (IES 2018)*. Atlantis Press. (DOI: 10.2991/ies-18.2019.9).

Maalej, W. & Nabil, H. 2015. 'Bug report, feature request, or simply praise? On automatically classifying app reviews'. *2015 IEEE 23rd International Requirements Engineering Conference, RE 2015 - Proceedings*. P. 116–125. ISBN 9781467369053. (DOI: 10.1109/RE.2015.7320414).

Mancosu, M. & Vegetti, F. 2020. 'What You Can Scrape and What Is Right to Scrape: A Proposal for a Tool to Collect Public Facebook Data'. *Social Media and Society*. 6(3). ISSN 20563051. (DOI: 10.1177/2056305120940703).

Markoff, J. 2006. *What the dormouse said: how the sixties counterculture shaped the personal computer industry*. New York, NY, USA: Penguin Books. ISBN 0143036769.

Maslowska, E., Malthouse, E. C. & Bernritter, S. F. 2017. 'Too good to be true: The role of online reviews' features in probability to buy'. *International Journal of Advertising*. 36(1). P. 142–163. ISSN 02650487. (DOI: 10.1080/02650487.2016.1195622).

Mou, J. *et al.* 2017. 'Predicting buyers' repurchase intentions in cross-border E-commerce: A valence framework perspective'. *Proceedings of the 25th European Conference on Information Systems, ECIS 2017*. 2017. P. 2382–2394. ISBN 9780991556700.

Mou, J. *et al.* 2019. 'International buyers' repurchase intentions in a Chinese cross-border e-commerce platform'. *Internet Research*. 30(2). P. 403–437. ISSN 1066-2243. (DOI: 10.1108/INTR-06-2018-0259).

Pang, B. & Lee, L. 2004. 'A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts'. in *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics - ACL '04*. Morristown, NJ, USA: Association for Computational Linguistics. P. 271-es. ISSN 0009-4978. (DOI: 10.3115/1218955.1218990).

Pang, B. & Lee, L. 2008. 'Opinion Mining and Sentiment Analysis'. *Foundations and Trends® in Information Retrieval*. 2(1–2). P. 1–135. ISSN 1554-0669. (DOI: 10.1561/1500000011).

Pang, B., Lee, L. & Vaithyanathan, S. 2002. 'Thumbs up? Sentiment Classification using Machine Learning Techniques'. in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - EMNLP '02*. Morristown, NJ, USA: Association for Computational Linguistics. P. 79–86. (DOI: 10.3115/1118693.1118704).

Panichella, S. *et al.* 2015. 'How can I improve my app? Classifying user reviews for software maintenance and evolution'. *2015 IEEE 31st International Conference on Software Maintenance and Evolution, ICSME 2015 - Proceedings*. (2). P. 281–290. ISBN 9781467375320. (DOI: 10.1109/ICSM.2015.7332474).

Pricearchive.org. 2020. *About Pricearchive.org*. *pricearchive.org*. [Last accessed 1 Mar. 2021] Available at: https://www.pricearchive.org/about.

Wang, Yu, Wang, Yi & Lee, S. H. 2017. 'The effect of cross-border e-commerce on China's international trade: An empirical study based on transaction cost analysis'. *Sustainability (Switzerland)*. 9(11). P. 1–13. ISSN 20711050. (DOI: 10.3390/su9112028).

Womply Research. 2021. *Impact of online reviews on small business revenue*. *womply.com*. [Last accessed 1 Apr. 2021] Available at: https://www.womply.com/impact-of-online-reviews-on-small-business-revenue/.

Wu, J. *et al.* 2013. 'User reviews and uncertainty assessment: A two stage model of consumers' willingness-to-pay in online markets'. *Decision Support Systems*. 55(1). P. 175–185. ISSN 01679236. (DOI: 10.1016/j.dss.2013.01.017).

Xu, F. 2016. *Alibaba vs. Amazon: A business model comparison*. Louvan School of Management, Université catholique de Louvain. Available at: http://hdl.handle.net/2078.1/thesis:7258.

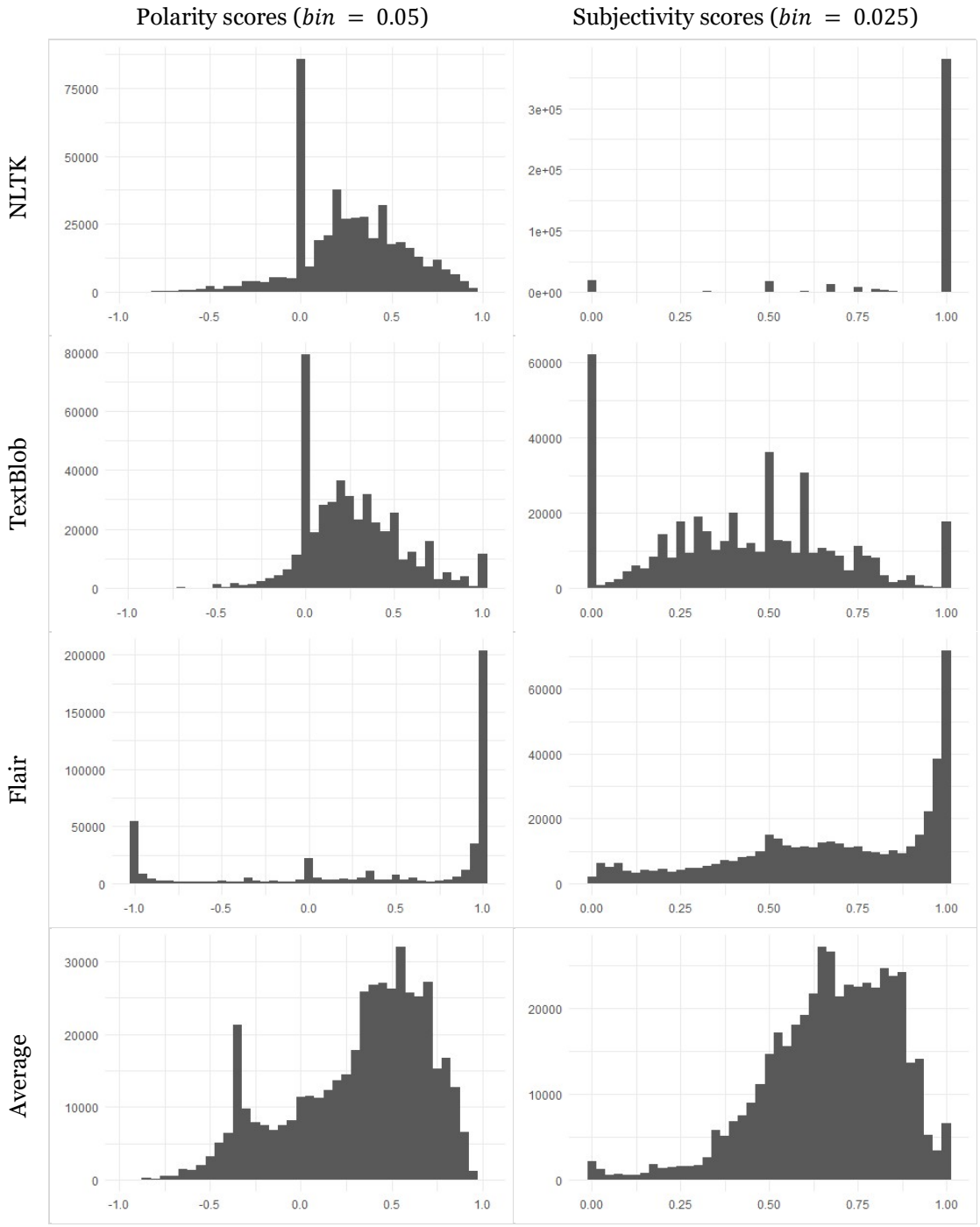
Zhang, Y. *et al.* 2014. 'Do users rate or review? Boost phrase-level sentiment labeling with review-level sentiment classification'. *SIGIR 2014 - Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*. (April 2016). P. 1027–1030. ISBN 9781450322591. (DOI: 10.1145/2600428.2609501).

Zhu, W., Mou, J. & Benyoucef, M. 2019. 'Exploring purchase intention in cross-border E-commerce: A three stage model'. *Journal of Retailing and Consumer Services*. 51(May). P. 320–330. ISSN 09696989. (DOI: 10.1016/j.jretconser.2019.07.004).

Appendices

Appendix 1: Sentiment analysis histograms

The following charts display the approximate distributions of review-level polarity and subjectivity scores using the three NLP libraries and the libraries' averages.



Appendix 2: Correlation matrices and VIF calculations

Appendix 2a: Correlation matrix for phone screen protectors ($n = 2{,}295$)

	$\ln(orders)$	$\ln(price)$	$\ln(age)$	$rating$	$sale$	SPD	$polarity$	$subjectiv.$
$\ln(orders)$	1							
$\ln(price)$	−0.0196	1						
$\ln(age)$	−0.0641	0.0369	1					
$rating$	0.1543	−0.2941	−0.1701	1				
$sale$	0.0172	−0.9472	−0.0473	0.344	1			
SPD	0.1516	0.4827	−0.0899	−0.0921	−0.4357	1		
$polarity$	0.3458	−0.0991	−0.0213	0.4306	0.1412	0.0437	1	
$subjectivity$	0.0107	0.0899	−0.0935	0.0152	−0.0236	0.0554	0.3478	1

Appendix 2b: Correlation matrix for phone charging cables ($n = 3{,}229$)

	$\ln(orders)$	$\ln(price)$	$\ln(age)$	$rating$	$sale$	SPD	$polarity$	$subjectiv.$
$\ln(orders)$	1							
$\ln(price)$	−0.5584	1						
$\ln(age)$	0.0587	−0.0131	1					
$rating$	0.0382	−0.0495	0.0405	1				
$sale$	0.5455	−0.8823	−0.0409	0.0834	1			
SPD	−0.0803	0.2511	0.0196	−0.0476	−0.1297	1		
$polarity$	0.2134	−0.0957	−0.0492	0.5222	0.122	−0.1126	1	
$subjectivity$	0.0618	0.0572	−0.1361	0.2532	−0.0009	−0.1373	0.5453	1

Appendix 2c: Correlation matrix for consumer electronics ($n = 2{,}795$)

	$\ln(orders)$	$\ln(price)$	$\ln(age)$	$rating$	$sale$	SPD	$polarity$	$subjectiv.$
$\ln(orders)$	1							
$\ln(price)$	−0.2024	1						
$\ln(age)$	−0.0167	−0.029	1					
$rating$	−0.0044	−0.0308	0.1082	1				
$sale$	0.3179	−0.3282	−0.1083	−0.0924	1			
SPD	0.0778	0.0815	−0.1298	−0.063	0.0769	1		
$polarity$	0.0431	0.1298	0.0254	0.4799	−0.0551	−0.0239	1	
$subjectivity$	−0.0053	0.2422	−0.057	0.1326	−0.0336	−0.0059	0.4904	1

Appendix 2d: VIF calculations for all regression models

Variable	Screen protectors		Phone charging cables		Consumer electronics	
	w/ *sale*	w/o *sale*	w/ *sale*	w/o *sale*	w/ *sale*	w/o *sale*
$\ln(price)$	10.897	1.4376	5.1519	1.1002	1.2247	1.0768
$\ln(age)$	1.0666	1.0660	1.0370	1.0252	1.0433	1.0331
$rating$	1.4591	1.4054	1.3903	1.3851	1.3599	1.3472
$sale$	10.610	n/a	4.8751	n/a	1.1637	n/a
SPD	1.3465	1.3403	1.1520	1.0943	1.0355	1.0265
$polarity$	1.4634	1.4631	1.8704	1.8704	1.7164	1.7159
$subjectivity$	1.2374	1.1974	1.5113	1.4984	1.4035	1.4004

Appendix 3: Heatmap of review-level polarities and ratings

Sentiment polarity	Numerical Rating					Total
	1 star	2 stars	3 stars	4 stars	5 stars	
0.9 to 1.0	0	0	4	56	3,858	3,918
0.8 to 0.9	1	2	20	379	23,178	23,580
0.7 to 0.8	11	7	101	1,068	40,862	42,049
0.6 to 0.7	11	29	111	933	46,262	47,346
0.5 to 0.6	23	23	166	1,176	58,784	60,172
0.4 to 0.5	41	31	201	1,249	51,008	52,530
0.3 to 0.4	124	98	306	1,523	46,922	48,973
0.2 to 0.3	106	60	292	1,089	27,810	29,357
0.1 to 0.2	111	98	349	1,211	23,235	25,004
0.0 to 0.1	209	153	546	1,567	20,393	22,868
−0.1 to 0.0	606	243	832	1,733	14,146	17,560
−0.2 to −0.1	898	362	950	1,609	10,410	14,229
−0.3 to −0.2	1,451	592	1,510	1,835	10,778	16,166
−0.4 to −0.3	6,802	1,716	3,823	3,231	14,657	30,229
−0.5 to −0.4	2,223	764	1,634	1,286	3,977	9,884
−0.6 to −0.5	998	334	613	456	1,451	3,852
−0.7 to −0.6	735	231	368	298	857	2,489
−0.8 to −0.7	280	77	126	86	259	828
−0.9 to −0.8	173	37	46	23	48	327
−1 to −0.9	6	1	1	4	2	14
Total	14,809	4,858	11,999	20,812	398,897	451,375